Master's Degree in Computational Mathematics

Master's Thesis

Probability distribution of the classical implication intensity seen as a random variable in Statistical Implicative Analysis

Author:
Paloma Marín Martínez

Tutor:
Pablo Gregori Huerta

Academic course 2016/2017

This is to certify that the present document entitled "Probability distribution of the classical implication intensity seen as a random variable in Statistical Implicative Analysis" constitutes the Master thesis of the student Paloma Marín Martínez, and is the result of her own work under my supervision.

Pablo Gregori Huerta
Associate Professor of Statistics
Universitat Jaume I de Castellón

Abstract

Starting from mathematical didactic situations, the method of Statistical Implicative Analysis is developed in correspondence with the problems encountered and the issues raised. Its main objective is to structure data, interrelating subjects and variables, extracting inductive rules between variables and, from the contingency of these rules, the explanation and consequently a certain forecast in different areas: psychology, sociology, biology, etc. This is why the concepts of intensity of implication, class cohesion, implication-inclusion, meaning of hierarchical levels, contribution of supplementary variables, etc. were created. In this work, most of the fundamental concepts of Statistical Implicative Analysis (SIA) are offered. We also study the behavior of the classical Gras implication index as a random variable, when applied to a couple of Bernoulli variables (X, Y), independent or not.

Keywords

- Statistical Implicative Analysis (SIA)
- Classical Gras implication index
- Binary variable
- Rule

Contents

Chapter 1

Introduction

The knowledge of humanity is constituted mainly according to two components: facts and rules between facts or between the same rules.

Unlike in mathematics, where every rule (theorem) has no exception, rules in humanistic sciences, more generally in so-called "soft" sciences, are acceptable and therefore applicable, while the number of counterexamples remains *admissible*. The problem, in the analysis of the data, is to establish a numerical criterion, to define the notion of confidence level adjustable to the level of demand of the user of the rule.

Before the reflection "if one exercise is more complex than another, then each student who passes the first will also pass the second", anyone knows that this situation usually presents exceptions. This motivated Régis Gras to model the extraction and representation of imprecise (or partial) inference rules between binary variables (or attributes or characters) that describe a population of individuals (subjects or objects). It is about discovering non-symmetric inductive rules to model relations of the type *if a, then almost b*.

To mathematize it, as with the method of measuring the similarity of I.C. Lerman, Gras defined the measure of confirmatory quality of the implicative relationship $a \rightarrow b$, from the implausibility of the occurrence in the data of the number of cases that invalidate it, that is, for which a is verified, but not b.

This fact was the origin of the Statistical Implicative Analysis in the 70's. Nowadays, there is an International Group of Implicative Statistic Analysis that, since 2000, has been holding International Conferences in different locations (France, Brazil, Italy, Spain, and Tunis) with the aim of disseminating these novel techniques of multivariate statistical analysis in different languages.

Chapter 2

The three tools of Statistical Implicative Analysis

The origin of the Statistical Implicative Analysis (SIA) was measuring the strength of the rules by using probabilities. The probabilistic measure of distances (or similarity indeed) among binary variables was done first by Lerman in 1970 ([4]), within cluster analysis. Then this symmetric similarity (cluster) analysis was adopted in the package of SIA. After, the evolution of the main pillar of SIA (the implication analysis), lead to a second hierarchical method, involving intensity of implications. It was called cohesion analysis and it constitutes the third pillar of SIA.

Let us consider a set I with n individuals, $I = \{x, y, z, \ldots\}$, and a set V with p properties, $V = \{a, b, c, d, \ldots\}$. If an individual $x \in I$ has property a, then $a(x) = 1$, otherwise $a(x) = 0$. Thus, we have as many binary variables as properties. We denote $A = \{x \in I : a(x) = 1\}$ and n_a as the number of individuals that have property a, i.e. the number of successes of variable a. Table 2.1 shows the contingency table of two of those dichotomous variables, a and b.

Example 1 *In order to exemplify all the theory in this work, we will use data from the research developed by Pitarch in 2002 ([6]) on students of Educación Secundaria Obligatoria (ESO). Students were asked if they liked some types of music. We will show only the results associated to some of the variables used in the research (see Table 2.2).*

	b	$\overline{b}$	
a	n_{ab}	$n_{a\overline{b}}$	n_a
$\overline{a}$	$n_{\overline{a}b}$	$n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

Table 2.1: Contingency table of dichotomous variables a and b.

	HIP	JAZ	HEA	REG	PUN
$s1$	1	0	0	0	1
$s2$	1	1	1	1	1
$s3$	1	0	0	0	1
$s4$	0	1	1	1	0
$s5$	1	1	1	1	1
$s6$	0	0	0	0	0
$s7$	1	1	1	1	1
$s8$	0	1	0	0	0
$s9$	1	0	0	1	1
$s10$	0	0	0	0	0
$s11$	0	1	1	1	1
$s12$	1	1	1	1	1
$s13$	1	0	1	0	1
$s14$	0	0	0	0	0
$s15$	0	1	0	0	0
$s16$	1	1	1	1	1
$s17$	1	0	0	1	0
$s18$	0	0	0	0	0
$s19$	1	1	1	1	1
$s20$	0	0	1	0	0

Table 2.2: Music data matrix.

2.1 Similarity

Definition 1 *The similarity index between two observed variables a and b, having n_{ab} copresences, is defined as the probability that randomness gives place to so many or fewer copresences as observed in the sample, i.e.*

$$s(a,b) := P(N_{XY} \leq n_{ab})$$

where N_{XY} indicates the random variable that counts the number of observed copresences between two variables, following a certain random process.

According to the sampling process, the random variable N_{XY} may have different probability distributions.

Definition 2 *A model for N_{XY} consists of creating two independent Bernoulli's trials samples with as many individuals as in the original sample, and with probability equal to the proportion of successes in each variable (n_a/n and n_b/n respectively). The appearance of copresences becomes a Bernoulli trial of success probability $n_a n_b/n^2$ (because of the independence). This implies that the variable N_{XY}, which counts the number of copresences after n trials, follows the binomial distribution $B(n = n, p = n_a n_b/n^2)$. It is called the Binomial model.*

Another model is to assume that the sample size was random, and followed the Poisson law of mean exactly n (the observed sample size). Then the previous process is repeated with the

(a,b)	λ	n_{ab}	$s(a,b)$
(HIP, JAZ)	$11 \cdot 10/20$	6	0.5844148
(HIP, HEA)	$11 \cdot 10/20$	7	0.7387844
(HIP, REG)	$11 \cdot 10/20$	8	0.856789
(HIP, PUN)	$11 \cdot 11/20$	10	0.9458524
(JAZ, HEA)	$10 \cdot 10/20$	8	0.9101438
(JAZ, REG)	$10 \cdot 10/20$	8	0.9101438
(JAZ, PUN)	$10 \cdot 11/20$	7	0.7387844
(HEA, REG)	$10 \cdot 10/20$	8	0.9101438
(HEA, PUN)	$10 \cdot 11/20$	8	0.856789
(REG, PUN)	$10 \cdot 11/20$	8	0.856789

Table 2.3: Similarity indices.

random sample size, and respective success probabilities proportional to n_a and n_b. Thus, two new realisations of the numbers of successes are created, N_a and N_b, and proceeded as in the previous binomial model. The development of the probability function of the variable N_{XY} leads to show that it follows the Poisson model $Po(\lambda = n_a n_b/n)$.

Example 2 *Let us calculate the similarity index of each pair of variables considered in the music data of Example 1.*

The calculation depends on the law assumed. So we will work with a Poisson model, and we will approximate it to the Normal distribution. If $N_{XY} \sim Po(\lambda)$ where $\lambda = \frac{n_a n_b}{n}$, we approximate it to $N_{XY} \approx N(\lambda, \sqrt{\lambda})$.

Then, for the pair (HIP,JAZ) we have $\lambda = \frac{n_{HIP} n_{JAZ}}{n} = \frac{11 \cdot 10}{20}$, and using R we get the similarity index $s(HIP, JAZ) = P(N_{XY} \leq 6) = 0.5844148$.

The indices of the rest of pairs are calculated in the same way and are shown in Table 2.3.

If we consider classes of grouped variables, we can also calculate the similarity between two of those classes.

Definition 3 *Let C_1 and C_2 be two nonempty disjoint subsets of the set of variables V. The similarity index between the classes of variables C_1 and C_2 is defined as*

$$s(C_1, C_2) := \max_{\substack{a \in C_1 \\ b \in C_2}} s(a,b)^{Card(C_1) \times Card(C_2)}$$

where Card indicates the cardinal of the referred set between parenthesis.

The algorithm that gives rise to the hierarchical tree of similarity begins by considering all p variables of V in isolation at level 0 as classes of variables. At the next level, the two classes

	HIP	JAZ	HEA	REG	PUN
HIP		0.5844148	0.7387844	0.856789	0.9458524
JAZ	0.5844148		0.9101438	0.9101438	0.7387844
HEA	0.7387844	0.9101438		0.9101438	0.856789
REG	0.856789	0.9101438	0.9101438		0.856789
PUN	0.9458524	0.7387844	0.856789	0.856789	

Table 2.4: Similarity matrix at level zero.

	(HIP,PUN)	JAZ	HEA	REG
(HIP,PUN)		0.5458024	0.7340874	0.7340874
JAZ	0.5458024		0.9101438	0.9101438
HEA	0.7340874	0.9101438		0.9101438
REG	0.7340874	0.9101438	0.9101438	

Table 2.5: Similarity matrix at level one.

with the highest index of similarity are grouped to form a new class and the rest of classes remain the same. And so on. For some examples of hierarchical trees, see [5].

Example 3 *Following with Example 1, we build the table of similarity at level zero with the indices of similiraty calculated before (see Table 2.4).*

To build the table of similarity at level one, we group variables HIP and PUN, since they have the highest index of similarity. The indices of similarity of the pairs of isolated variables remain the same, but now we have to calculate the indices where the class (HIP,PUN) is involved.

$$\begin{aligned} s((HIP,PUN),JAZ) &= \max\{s(HIP,JAZ),s(PUN,JAZ)\}^2 = \\ &= \max\{0.5844148,0.7387844\}^2 = 0.7387844^2 = 0.5458024 \end{aligned}$$

The indices of the rest of pairs are calculated in the same way and are shown in Table 2.5.

There are three pairs of variables with the highest index of similarity: (JAZ, HEA), (JAZ, REG), and (HEA, REG). So we can group any of them in the next level of the hierarchical tree. In this case, we choose the pair (JAZ, HEA) because it is the first of the list. Now we need to calculate the indices where the class (JAZ, HEA) is involved, the rest of indices remain the same.

$$\begin{aligned} &s((HIP,PUN),(JAZ,HEA)) = \\ &= \max\{s(HIP,JAZ),s(HIP,HEA),s(PUN,JAZ),s(PUN,HEA)\}^4 = \\ &= \max\{0.5844148,0.7387844,0.7387844,0.856789\}^4 = 0.856789^4 = 0.5388843 \end{aligned}$$

All indices are shown in the matrix of similarity at level two (Table 2.6).

	(HIP,PUN)	**(JAZ,HEA)**	**REG**
(HIP,PUN)		0.5388843	0.7340874
(JAZ,HEA)	0.5388843		0.8283617
REG	0.7340874	0.8283617	

Table 2.6: Similarity matrix at level two.

	(HIP,PUN)	**(JAZ,HEA,REG)**
(HIP,PUN)		0.3955882
(JAZ,HEA,REG)	0.3955882	

Table 2.7: Similarity matrix at level three.

Since the highest index is the index of similarity of (JAZ,HEA) and REG, we will group them in a new class at the next level. So, we calculate the new index.

$s((HIP, PUN), (JAZ, HEA, REG)) = \max\{s(HIP, JAZ), s(HIP, HEA), s(HIP, REG), s(PUN, JAZ), s(PUN, HEA), s(PUN, REG)\}^6 = \max\{0.5844148, 0.7387844, 0.856789, 0.7387844, 0.856789, 0.856789\}^6 = 0.856789^6 = 0.3955882$

You can see the matrix of similarity at level three at Table 2.7.

And at Figure 2.1 you can see the hierarchical tree, based on all these tables.

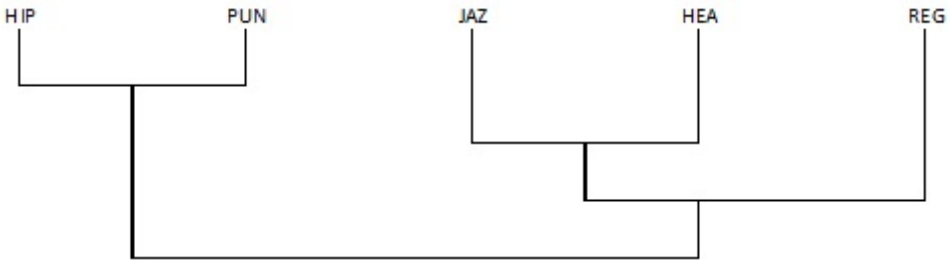

Figure 2.1: Hierarchical tree of similarity.

After the tree (also called dendrogram) is represented, a number of groups can be chosen, and the resulting groups of variables arise. For instance, if we want to form two groups, we split the last level and get $\{HIP, PUN\}$ and $\{JAZ, HEA, REG\}$. In case we want to consider three groups, we split the penultimate level, and get the groups $\{HIP, PUN\}$, $\{JAZ, HEA\}$ and $\{REG\}$. Music types in the same group would be considered as similar, in the sense that they all are liked together (or not) by enough members of the sample.

The following definition has the intention of highlight some specific levels of the similarity tree. The philosophy behind is the following: at each new level of aggregation, we can check every couple of variables, and see whether they have been gathered (at this same or another

earlier level) in a class of variables, or not. Ideally, couples of variables holding a large similarity index, should have been gathered earlier, but it is not always the case. When the present level of aggregation shows that gathered variables have generally the largest similarity index, then this level is interesting, compared to other levels.

Definition 4 *The preorder induced by the application S (similarity) over $V \times V$ is called the initial and global preorder Ω over $V \times V$.*

$$G_S(\Omega) = \{((a,b);(c,d)) : \ s(a,b) < s(c,d)\}$$

At each level i of the hierarchical tree, the same set of $p(p-1)/2$ pairs of different variables of V can be partitioned into two subsets, R_i and S_i. We say that there is a division Π_i.

- R_i, the set of pairs (a,b) of distinct variables that, at that level i, belong to the same class.
- S_i, the remaining pairs of variables.

In [3], a very easy and illustrative example of how these two sets work can be found.

Property 1 *(Gras and Kuntz, 2008) Under complete randomness (uniform distribution) in the set of possible initial preorders in V, the random variable that counts, for each level i, the number of pairs of $S_i \times R_i$ that respect the (random) preorder, follows a law with expected value $s_i \times r_i$ and variance $\frac{1}{12} s_i \times r_i \times (s_i + r_i + 1)$, where $s_i := Card(S_i)$ and $r_i := Card(R_i)$.*

Definition 5 *For each level $i = 1, 2, \ldots, p$, the $S(\Omega, i)$ index is defined as the typification of the number of $S_i \times R_i$ pairs through the mean and standard deviation mentioned in Property 1:*

$$S(\Omega, i) = \frac{Card[G(\Omega) \cap [S_i \times R_i]] - \frac{1}{2} s_i r_i}{\sqrt{\frac{s_i r_i (s_i + r_i + 1)}{12}}}$$

As in Lerman's analysis of similarities, this index serves as *global statistic of the levels.*

Definition 6 *We call significant level to every level that corresponds to a local maximum of $S(\Omega, k)$ during the construction of the hierarchy. In this case, we say that the division Π_k is in partial accordance with Ω. If, furthermore, $G(\Omega) \cap [S_k \times R_k] = [S_k \times R_k]$, we say that the division Π_k is in total accordance with Ω.*

Example 4 *Following with Example 1, we will see how these significant levels are determined.*

We denote:

- $t = Card(V)$, $I_k = \{j |\ a_j \in R_k\}$
- a_i *the elements of the initial preorder* Ω, *with* $i = 1, \ldots, 2 \cdot \binom{t}{2}$
- $P_l = \{a_j \in \Omega |\ s(a_j) = s_l\}$, *where* s_l *is any value that the similarity index may have, and* $l = 1, \ldots, d$, *where* d *is the amount of different values that this similarity index has*
- $f_i^k = Card(a_j |\ j \notin I_k,\ j < i,\ s(a_j) = s(a_i))$

From the similarity indices that are shown at Table 2.3, the initial and global preorder Ω *over* $V \times V$ *can be determined. We will show it using five columns, each of them has the pairs of variables with same similarity index:*

$$\left\{ \begin{bmatrix} (HIP, JAZ) \\ (JAZ, HIP) \end{bmatrix} < \begin{bmatrix} (HIP, HEA) \\ (HEA, HIP) \\ (JAZ, PUN) \\ (PUN, JAZ) \end{bmatrix} < \begin{bmatrix} (HIP, REG) \\ (REG, HIP) \\ (HEA, PUN) \\ (PUN, HEA) \\ (REG, PUN) \\ (PUN, REG) \end{bmatrix} < \right.$$
$$\left. < \begin{bmatrix} (JAZ, HEA) \\ (HEA, JAZ) \\ (JAZ, REG) \\ (REG, JAZ) \\ (HEA, REG) \\ (REG, HEA) \end{bmatrix} < \begin{bmatrix} (HIP, PUN) \\ (PUN, HIP) \end{bmatrix} \right\}$$

Then

$$G_S(\Omega) = \left\{ \begin{array}{l} ((HIP, JAZ), (HIP, HEA)); \ldots; ((HIP, JAZ), (PUN, HIP)); \\ ((JAZ, HIP), (HIP, HEA)); \ldots; ((JAZ, HIP), (PUN, HIP)); \\ \ldots; \\ ((REG, HEA), (HIP, PUN)); ((REG, HEA), (PUN, HIP)) \end{array} \right\}.$$

Now we will calculate its cardinal. Let $m = Card(\Omega) = 2 \cdot \binom{t}{2} = 2 \cdot \binom{5}{2} = 20$, *then*

$$\begin{aligned} Card(G_S(\Omega)) &= \tfrac{m \cdot (m-1)}{2} - \textstyle\sum_{l=1}^{d} \tfrac{Card(P_l) \cdot (Card(P_l)-1)}{2} = \\ &= \tfrac{20 \cdot 19}{2} - \left(\tfrac{2 \cdot 1}{2} + \tfrac{4 \cdot 3}{2} + \tfrac{6 \cdot 5}{2} + \tfrac{6 \cdot 5}{2} + \tfrac{2 \cdot 1}{2}\right) = \\ &= 190 - (1 + 6 + 15 + 15 + 1) = \\ &= 190 - 38 = \\ &= 152 \end{aligned}$$

The cardinal $Card[G(\Omega) \cap [S_k \times R_k]]$ *can be calculated by counting in* Ω *the pairs that belong to* $G(\Omega) \cap [S_k \times R_k]$, *which is achieved by suming all the elements that are on the left of the column that contains the pair that joins the* k*th level and that are not in the set* R_k.

***Level 1 of the hierarchy**: Only variables HIP and PUN are joined. Thus*

$R_1 = \{(HIP, PUN)\}$

$S_1 = \{(HIP, JAZ), (JAZ, HIP), (HIP, HEA), (HEA, HIP), (JAZ, PUN),$
$(PUN, JAZ), (HIP, REG), (REG, HIP), (HEA, PUN), (PUN, HEA), (REG, PUN),$
$(PUN, REG), (JAZ, HEA), (HEA, JAZ), (JAZ, REG), (REG, JAZ), (HEA, REG),$
$(REG, HEA), (PUN, HIP)\}$

$Card(S_1) = 19$, $Card(R_1) = 1$, $I_1 = \{19\}$, $f_1^1 = 1$, *and* $Card[G(\Omega) \cap [S_1 \times R_1]] = 18$.

$S(\Omega, 1) = \frac{Card[G(\Omega) \cap [S_1 \times R_1]] - \frac{1}{2} s_1 r_1}{\sqrt{\frac{s_1 r_1 (s_1 + r_1 + 1)}{12}}} = \frac{18 - \frac{1}{2} \cdot 19 \cdot 1}{\sqrt{\frac{19 \cdot 1 (19+1+1)}{12}}} = \frac{18 - 19/2}{\sqrt{19 \cdot 21/12}} = 1.474087.$

***Level 2 of the hierarchy**: The pair (JAZ, HEA) joines the assembled.*

$R_2 = \{(HIP, PUN), (JAZ, HEA)\}$

$S_2 = \{(HIP, JAZ), (JAZ, HIP), (HIP, HEA), (HEA, HIP), (JAZ, PUN),$
$(PUN, JAZ), (HIP, REG), (REG, HIP), (HEA, PUN), (PUN, HEA), (REG, PUN),$
$(PUN, REG), (HEA, JAZ), (JAZ, REG), (REG, JAZ), (HEA, REG), (REG, HEA),$
$(PUN, HIP)\}$

$Card(S_2) = 18$, $Card(R_2) = 2$, $I_2 = \{19, 13\}$, $f_1^2 = 1$, $f_2^2 = 5$, *and* $Card[G(\Omega) \cap [S_2 \times R_2]] = 29$.

$S(\Omega, 2) = \frac{29 - \frac{1}{2} \cdot 18 \cdot 2}{\sqrt{\frac{18 \cdot 2 (18+2+1)}{12}}} = \frac{29 - 18}{\sqrt{18 \cdot 21/6}} = 1.385870.$

***Level 3 of the hierarchy**: The class $((JAZ, HEA), REG)$ is formed, and then*

$R_3 = \{(HIP, PUN), (JAZ, HEA), (JAZ, REG), (HEA, REG)\}$

$S_3 = \{(HIP, JAZ), (JAZ, HIP), (HIP, HEA), (HEA, HIP), (JAZ, PUN),$
$(PUN, JAZ), (HIP, REG), (REG, HIP), (HEA, PUN), (PUN, HEA), (REG, PUN),$
$(PUN, REG), (HEA, JAZ), (REG, JAZ), (REG, HEA), (PUN, HIP)\}$

$Card(S_3) = 16$, $Card(R_3) = 4$, $I_3 = \{19, 13, 15, 17\}$, $f_1^3 = 1$, $f_2^3 = 3$, $f_3^3 = 3$, $f_4^3 = 3$, *and* $Card[G(\Omega) \cap [S_3 \times R_3]] = 51$.

$S(\Omega, 3) = \frac{51 - \frac{1}{2} \cdot 16 \cdot 4}{\sqrt{\frac{16 \cdot 4 (16+4+1)}{12}}} = 1.795331.$

***Level 4 of the hierarchy**: At this level, the class $((HIP, PUN), ((JAZ, HEA), REG))$ is formed. Thus*

$R_4 = \{(HIP, PUN), (JAZ, HEA), (JAZ, REG), (HEA, REG), (HIP, JAZ),$
$(HIP, HEA), (HIP, REG), (PUN, JAZ), (PUN, HEA), (PUN, REG)\}$

Level	s_k	r_k	Card$[G(\Omega) \cap [S_k \times R_k]]$	$S(\Omega, k)$
1	19	1	18	1.474087
2	18	2	29	1.385870
3	16	4	51	1.795331
4	10	10	38	-0.907115

Table 2.8: Values of the index $S(\Omega, k)$.

$S_4 = \{(JAZ, HIP), (HEA, HIP), (JAZ, PUN), (REG, HIP), (HEA, PUN), (REG, PUN), (HEA, JAZ), (REG, JAZ), (REG, HEA), (PUN, HIP)\}$

$Card(S_4) = 10$, $Card(R_4) = 10$, $I_4 = \{19, 13, 15, 17, 1, 3, 7, 6, 10, 12\}$, $f_1^4 = 1$, $f_2^4 = 3$, $f_3^4 = 3$, $f_4^4 = 3$, $f_5^4 = 1$, $f_6^4 = 2$, $f_7^4 = 3$, $f_8^4 = 2$, $f_9^4 = 3$, $f_{10}^4 = 3$, *and* $Card[G(\Omega) \cap [S_4 \times R_4]] = 38$.

$$S(\Omega, 4) = \frac{38 - \frac{1}{2} \cdot 10 \cdot 10}{\sqrt{\frac{10 \cdot 10(10+10+1)}{12}}} = -0.907115.$$

At Table 2.8 we sum up all obtained values. And Figure 2.2 shows the plot of the index $S(\Omega, k)$.

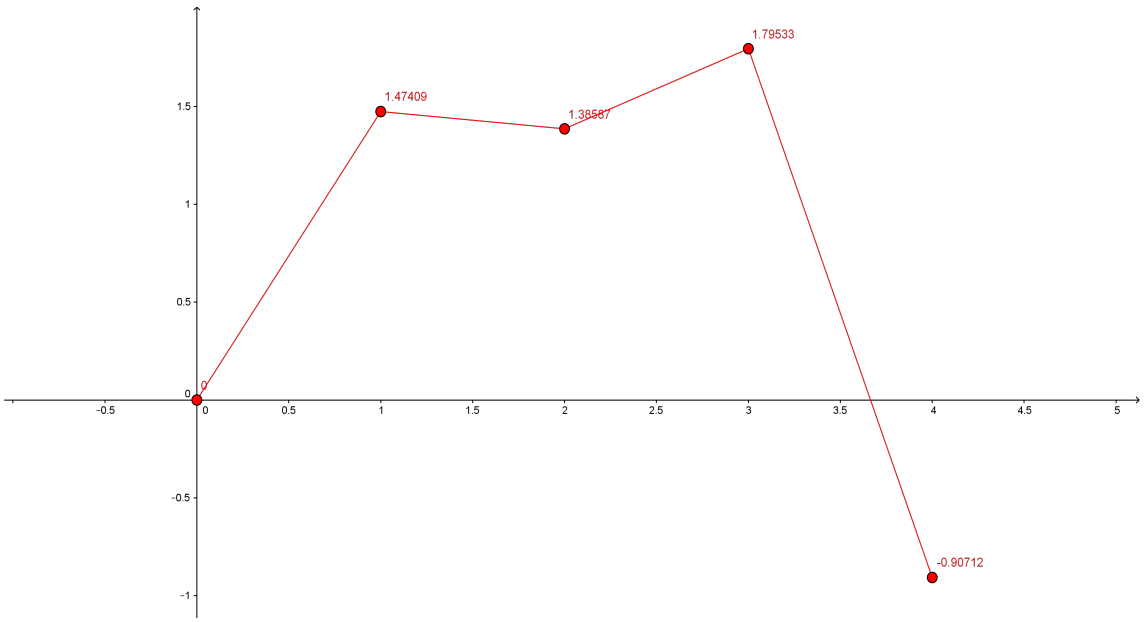

Figure 2.2: Plot of values of $S(\Omega, k)$.

According to Definition 6, the significant levels are levels 1 and 3, since they are local maximums of $S(\Omega, k)$ *(see Figure 2.2).*

2.2 Implication

Let us suppose that we select randomly, from I, two subsets X and Y with n_a and n_b elements respectively. Let $N_{X\overline{Y}} = \text{Card}(X \cap \overline{Y})$ the random variable associated to the number of counterexamples in this random model.

Definition 7 *The rule $a \to b$ is admissible at confidence level $1-\alpha$ if*

$$P(N_{X\overline{Y}} \leq n_{a\overline{b}}) \leq \alpha$$

where $n_{a\overline{b}} = Card(A \cap \overline{B})$ is the number of counterexamples of the rule $a \to b$ observed in the sample.

In a parallel way, as in the similarity analysis, the random variable $N_{X\overline{Y}}$ follows the binomial model B($n = n$,$p = n_a n_{\overline{b}}/n^2$) when the sample size is fixed previously, and the Poisson law of parameter $\frac{n_a n_{\overline{b}}}{n}$ when the sample size is assumed to be random, following the Poisson law of mean n.

When $n_{\overline{b}} \neq 0$, that Poisson variable can be reduced and centered. In the experimental realisation, the observed value of that reduced and centered variable estimates the difference between the contingency and the value that it had taken in case of independence between a and b. That estimation is defined as the implication index.

Definition 8 *The implication index of the rule $a \to b$ is defined as:*

$$q(a,\overline{b}) = \frac{n_{a\overline{b}} - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}},$$

number chosen as indicator of the no implication of a over b.

The implication intensity, quality of the admissibility of $a \to b$, for $n_a \leq n_b$ and $n_b \neq n$, is defined as follows.

Definition 9 *The implication intensity of the rule $a \to b$, which measures the inductive quality of a over b is:*

$$\varphi(a,b) = \begin{cases} 1 - P(N_{X\overline{Y}} \leq n_{a\overline{b}}) & \text{if } n_b \neq n \\ 0 & \text{if } n_b = n \end{cases}$$

Therefore, the definition of statistic implication is the following.

Definition 10 *The implication $a \to b$ is admissible at confidence level $1-\alpha$ if and only if $\varphi(a,b) \geq 1-\alpha$.*

$a \to b$	p	$n_{a\overline{b}}$	$\varphi(a,b)$
$HIP \to JAZ$	$11 \cdot 10/20^2$	5	0.4846112
$JAZ \to HIP$	$10 \cdot 9/20^2$	4	0.4798163
$HIP \to HEA$	$11 \cdot 10/20^2$	4	0.6805756
$HEA \to HIP$	$10 \cdot 9/20^2$	2	0.690122
$HIP \to REG$	$11 \cdot 10/20^2$	3	0.8420236
$REG \to HIP$	$10 \cdot 9/20^2$	2	0.8605659
$HIP \to PUN$	$11 \cdot 9/20^2$	1	0.9743144
$PUN \to HIP$	$11 \cdot 9/20^2$	1	0.9743144
$JAZ \to HEA$	$10 \cdot 10/20^2$	2	0.9087396
$HEA \to JAZ$	$10 \cdot 10/20^2$	1	0.9756874
$JAZ \to REG$	$10 \cdot 10/20^2$	2	0.9087396
$REG \to JAZ$	$10 \cdot 10/20^2$	2	0.9087396
$JAZ \to PUN$	$10 \cdot 9/20^2$	3	0.690122
$PUN \to JAZ$	$11 \cdot 10/20^2$	4	0.6805756
$HEA \to REG$	$10 \cdot 10/20^2$	2	0.9087396
$REG \to HEA$	$10 \cdot 10/20^2$	2	0.9087396
$HEA \to PUN$	$10 \cdot 9/20^2$	2	0.8605659
$PUN \to HEA$	$11 \cdot 10/20^2$	3	0.8420236
$REG \to PUN$	$10 \cdot 9/20^2$	2	0.8605659
$PUN \to REG$	$11 \cdot 10/20^2$	3	0.8420236

Table 2.9: Implication intensities. Values exceeding the confidence level 0.85 are highlighted in red.

Example 5 *Following with Example 1, we will compute the implication intensity of the rules formed by the five variables of music data. But now we will work with a Binomial model. So* $N_{X\overline{Y}} \sim Bin(n,p)$ *where* $p = \frac{n_a n_{\overline{b}}}{n^2}$.

Then, for the rule $HIP \to JAZ$ *we have* $n_{JAZ} = 10 \neq 20 = n$. *So we calculate* $p = \frac{n_{HIP} n_{\overline{JAZ}}}{n^2} = \frac{11 \cdot 10}{20^2}$, *and using R we get the implication intensity* $\varphi(HIP, JAZ) = 1 - P(N_{X\overline{Y}} \leq 5) = 1 - 0.5153888 = 0.4846112$.

The intensities of the rest of rules are calculated in the same way and are shown in Table 2.9.

It is known that for large samples, the implication intensity gets values that do not discriminate ([1]), so a new index of implication-inclusion was defined (also called entropic version of the implication, as opposed to the classical version that we have presented).

First, we consider the conditional entropies:

$$H(b|a) = -\frac{n_{ab}}{n_a} \log_2 \left(\frac{n_{ab}}{n_a} \right) - \frac{n_{a\overline{b}}}{n_a} \log_2 \left(\frac{n_{a\overline{b}}}{n_a} \right)$$

$$H(\overline{a}|\overline{b}) = -\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} \log_2 \left(\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} \right) - \frac{n_{a\overline{b}}}{n_{\overline{b}}} \log_2 \left(\frac{n_{a\overline{b}}}{n_{\overline{b}}} \right)$$

which measure the uncertainty of observing b when observed a, and respectively, observing $\overline{a}$ when observed $\overline{b}$. Second, it is considered that it makes no sense to think of an inclusion when 50% of the elements of the hypothetical subset A do not belong to the hypothetical set that contains it, B. That is why the conditional entropies are truncated by defining:

$$h(b|a) = \begin{cases} H(b|a), & \text{if } \frac{n_{a\overline{b}}}{n_a} \leq 0.5 \\ 1, & \text{if } \frac{n_{a\overline{b}}}{n_a} > 0.5 \end{cases}$$

$$h(\overline{a}|\overline{b}) = \begin{cases} H(\overline{a}|\overline{b}), & \text{if } \frac{n_{a\overline{b}}}{n_{\overline{b}}} \leq 0.5 \\ 1, & \text{if } \frac{n_{a\overline{b}}}{n_{\overline{b}}} > 0.5 \end{cases}$$

Definition 11 *We define the index of the inclusion $A \subset B$ as*

$$i(a,b) = ((1 - h(b|a))^2(1 - h(\overline{a}|\overline{b}))^2)^{1/4}$$

Now we define the new implication intensity index, which does not have the drawback that the classical index has.

Definition 12 *We define the index of implication-inclusion of the rule $a \to b$ as*

$$\Psi(a,b) = (\varphi(a,b)^2 i(a,b)^2)^{1/4}$$

If we choose a threshold for the index, for example 0.95, a natural binary relationship appears (reflexive, but neither symmetric nor transitive) between variables: aRb if and only if $\Psi(a,b) \geq 0.95$. This binary relationship can be represented graphically.

Definition 13 *Given a threshold, we call implicative graph of the sample (for the threshold) to the directed graph, whose vertices are the variables of the sample, associated to the binary relationship R, where aRb if there is an arrow from a to b with index not lower than the threshold.*

Example 6 *Following with Example 1, if we look at Table 2.9 we see that all rules coloured in red are admissible at confidence level* 0.85*. Figure 2.3 shows the implicative graph for this sample.*

The interpretation of every rule is the same. For instance, as $REG \to HIP$ is the admissible at confidence level 0.85*, liking REG music improves significantly the chances of liking HIP music. We don't really know the chances, but we know that they improve significantly, for the level* $1 - 0.85$*, comparing the complete sample with the sample of people liking REG.*

At Table 2.10 we have calculated the index of implication-inclusion for each rule. As we can see, with this index, only two rules are admissible at confidence level 0.70 (coloured in red). We get the graph shown in Figure 2.4.

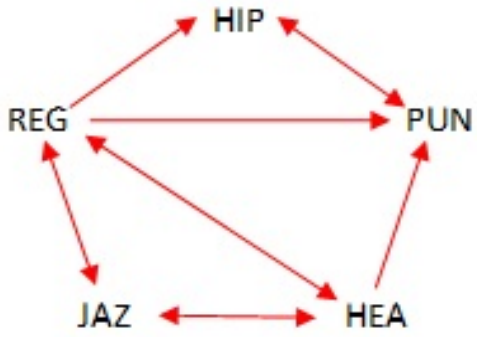

Figure 2.3: Implicative graph.

Figure 2.4: Implicative graph considering the index of implication-inclusion.

2.3 Cohesion

Definition 14 *R-rules are subsets of variables of V formed according to the following algorithm:*

- *An R-rule of degree 0 is any variable of V.*
- *An R-rule of degree 1 is any implication between two different rules of degree 0 (i.e. variables).*
- *By induction, for each $i > 1$, an implication $R_1 \to R_2$ of two R-rules of respective degrees g_1 and g_2 such that $g_1 + g_2 = i - 1$, and which involve two disjoint subsets of variables, form an R-rule of degree i.*

If we consider an experiment with an R-rule of degree 1, $a \to b$, the entropy is

$$E = -p \log_2 p - (1-p) \log_2 (1-p), \quad \text{with } p = \varphi(a,b)$$

Definition 15 *The cohesion of a rule $a \to b$ is:*

$$c(a,b) = \begin{cases} \sqrt{1-E^2} & \text{if } \varphi(a,b) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Note that the cohesion is truncated to 0 at the moment that the implication intensity does not reach the value 0.5.

$a \to b$	$\varphi(a,b)$	$i(a,b)$	$\Psi(a,b)$
$HIP \to JAZ$	0.4846112	0	0
$JAZ \to HIP$	0.4798163	0.0161008	0.0878944
$HIP \to HEA$	0.6805756	0.0397308	0.1644379
$HEA \to HIP$	0.690122	0.0984836	0.2607024
$HIP \to REG$	0.8420236	0.1354926	0.3377691
$REG \to HIP$	0.8605659	0.2560627	0.4694239
$HIP \to PUN$	0.9743144	0.5276601	0.7170124
$PUN \to HIP$	0.9743144	0.5276601	0.7170124
$JAZ \to HEA$	0.9087396	0.2780719	0.5026877
$HEA \to JAZ$	0.9756874	0.4102647	0.6326848
$JAZ \to REG$	0.9087396	0.2780719	0.5026877
$REG \to JAZ$	0.9087396	0.2780719	0.5026877
$JAZ \to PUN$	0.690122	0.0984836	0.2607024
$PUN \to JAZ$	0.6805756	0.0397308	0.1644379
$HEA \to REG$	0.9087396	0.2780719	0.5026877
$REG \to HEA$	0.9087396	0.2780719	0.5026877
$HEA \to PUN$	0.8605659	0.2560627	0.4694239
$PUN \to HEA$	0.8420236	0.1354926	0.3377691
$REG \to PUN$	0.8605659	0.2560627	0.4694239
$PUN \to REG$	0.8420236	0.1354926	0.3377691

Table 2.10: Indices of implication-inclusion. Values exceeding the level 0.7 are highlighted in red.

Definition 16 *The cohesion of a class of variables $R = \{a_1, \ldots, a_k\}$ is the geometric mean of the cohesions of the pairs of variables*

$$c(R) = \left\{ \prod_{\substack{i,j \\ j>i}} c(a_i, a_j) \right\}^{\frac{2}{k(k-1)}}$$

Definition 17 *The implication intensity of a class A over a class B is:*

$$\psi(A,B) = \left\{ \sup_{\substack{i=1,\ldots,r \\ j=1,\ldots,s}} \varphi(a_i, b_j) \right\}^{r \times s} [c(A)c(B)]^{\frac{1}{2}}$$

where $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$.

Example 7 *Following with Example 1, we will compute the index of cohesion of the rules formed by the five variables of music data.*

Then, for the rule $HIP \to JAZ$ we have $\varphi(HIP, JAZ) = 0.4846112 < 0.5$. So, applying the formula of Definition 15, we get $c(HIP, JAZ) = 0$.

	HIP	JAZ	HEA	REG	PUN
HIP		0	0.4280513	0.7770439	0.9850496
JAZ	0		0.8976726	0.8976726	0.449992
HEA	0.449992	0.9862913		0.8976726	0.8126480
REG	0.8126480	0.8976726	0.8976726		0.812648
PUN	0.9850496	0.4280513	0.7770439	0.7770439	

Table 2.11: Cohesion matrix at level zero.

	(HEA,JAZ)	HIP	REG	PUN
(HEA,JAZ)		0.4729926	0.8201277	0.73548
HIP	0.4599974		0.7770439	0.9850496
REG	0.8201277	0.8126480		0.812648
PUN	0.7041272	0.9850496	0.770439	

Table 2.12: Cohesion matrix at level one.

The indices of the rest of rules are calculated in the same way and are shown in Table 2.11.

To build the table of cohesion at level one, we group variables HEA and JAZ, since they have the highest index of cohesion. The indices of cohesion of the pairs of isolated variables remain the same, but now we have to calculate the indices where the class (HEA,JAZ) is involved. Applying the formula of Definition 17, we get:

$$\begin{aligned}\psi((HEA,JAZ),HIP) &= \sup\{\varphi(HEA,HIP),\varphi(JAZ,HIP)\}^2 \cdot [c(HEA,JAZ)\cdot c(HIP)]^{\frac{1}{2}} \\ &= \sup\{0.690122, 0.4798163\}^2 \cdot [0.9862913 \cdot 1]^{\frac{1}{2}} = 0.4729926\end{aligned}$$

The indices of the rest of pairs are calculated in the same way and are shown in Table 2.12.

The maximum cohesion is between variables HIP and PUN and PUN and HIP with a cohesion index, in both cases, equal to 0.9850496. So we group them in the next level of the hierarchical tree. Now we need to calculate the indices where the class (HIP,PUN) is involved, the rest of indices remain the same.

$$\begin{aligned}&\psi((HEA,JAZ),(HIP,PUN)) = \\ &= \sup\{\varphi(HEA,HIP),\varphi(HEA,PUN),\varphi(JAZ,HIP),\varphi(JAZ,PUN)\}^4 \\ &\cdot[c(HEA,JAZ)\cdot c(HIP,PUN)]^{\frac{1}{2}} = \\ &= \sup\{0.690122, 0.8605659, 0.4798163, 0.690122\}^4 \cdot [0.9862913 \cdot 0.9850496]^{\frac{1}{2}} = 0.5405902\end{aligned}$$

All indices are shown in the matrix of cohesion at level two (Table 2.13).

Since the highest index is the index of cohesion of (HEA,JAZ) and REG, we will group them in a new class at the next level. So, we calculate the new index.

	(HEA,JAZ)	(HIP,PUN)	REG
(HEA,JAZ)		0.5405902	0.8201277
(HIP,PUN)	0.4954829		0.7036838
REG	0.8201277	0.7350169	

Table 2.13: Cohesion matrix at level two.

	(HEA,JAZ,REG)	(HIP,PUN)
(HEA,JAZ,REG)		0.3879783
(HIP,PUN)	0.3404459	

Table 2.14: Cohesion matrix at level three.

First, we use the formula of Definition 16 to calculate the cohesion of the new class formed.

$$\begin{aligned} c(HEA, JAZ, REG) &= \{c(HEA, JAZ) \cdot c(HEA, REG) \cdot c(JAZ, REG)\}^{\frac{2}{6}} = \\ &= \{0.9862913 \cdot 0.8976726 \cdot 0.8976726)^{\frac{1}{3}} = 0.9262902 \end{aligned}$$

And now we calculate the implication intensity of (HEA,JAZ,REG) over (HIP,PUN).

$$\psi((HEA, JAZ, REG), (HIP, PUN)) = \sup\{\varphi(HEA, HIP), \varphi(HEA, PUN), \varphi(JAZ, HIP),$$
$$\varphi(JAZ, PUN), \varphi(REG, HIP), \varphi(REG, PUN)\}^6 \cdot [c(HEA, JAZ, REG) \cdot c(HIP, PUN)]^{\frac{1}{2}} =$$
$$= \sup\{0.690122, 0.8605659, 0.4798163, 0.690122, 0.8605659, 0.8605659\}^6$$
$$\cdot[0.9262902 \cdot 0.9850496]^{\frac{1}{2}} = 0.3879783$$

The other index is calculated in the same way. You can see the matrix of cohesion at level three at Table 2.14.

And at Figure 2.5 you can see the hierarchical tree, based on all these tables.

Figure 2.5: Hierarchical tree of cohesion

The theory of SIA introduces the concept of significant level of the cohesion tree, as it does it in the similarity tree, but we shall not cover it in the present work.

2.4 Typicality and contribution

The procedure to determine the typicality and contribution of the individuals to the classes that are formed in the similarity and cohesitive trees are the same. Formulas are different only in the fact that in the first case we use similarity index, and in the second case, the cohesion index between variables.

Definition 18 *It is defined as a typical subject, that which verifies all the implications that have the highest implication intensity in the formation of classes.*

The aim of the definition of typicality is to identify those individuals that are *typical* of the behavior of the population, that is, they comply with the rules with highest intensity of implication.

Definition 19 *Let i be a level of the hierarchy, and C the class formed at that level by bringing together the classes C_1 and C_2, present at level $i-1$. The pair (a,b) with $a \in C_1$ and $b \in C_2$ such that*

$$\varphi(a,b) \geq \varphi(j,k) \quad \forall j \in C_1 \quad \text{and} \quad \forall k \in C_2,$$

is called generic pair of the class C. The number $\varphi(a,b) = \varphi_i$ is called generic implication of C.

Definition 20 *Given that both C_1 and C_2 are classes gathered at any inferior level $g < i$, then we can consider the vector $(\varphi_1, \ldots, \varphi_i) \in [0,1]^i$ that is called implicative vector of the class C built at level i.*

By logical and semantic considerations, for each individual x in the sample, we denote by $\varphi_{x,i}$ the values:

- $\varphi_{x,i} = 1$ if $b(x) = 1$ (because x is not a counterexample of the rule).
- $\varphi_{x,i} = 0$ if $a(x) = 1$ and $b(x) = 0$ (because x is a counterexample of the rule).
- $\varphi_{x,i} = 0.5$ if $a(x) = b(x) = 0$ (because x does not present any of both attributes, and it must not count as a true example of the rule).

Using this, we define a distance between an individual and a formed class, through the expression:

$$d_1(x,C) := \left(\frac{1}{g} \sum_{i=1}^{g} \frac{(\varphi_i - \varphi_{x,i})^2}{1-\varphi_i} \right)^{1/2}$$

where g are all subclasses of C formed at previous levels (including C).

Definition 21 *We define the typicality of the individual x to the formation of class C of level i of the hierarchical tree as*

$$\gamma_1(x, C) := 1 - \frac{d_1(x, C)}{\max_{y \in I} d_1(y, C)}$$

Typicality of a supplementary variable (i.e., not included in the analysis) is the mean of the typicalities of the individuals marked by this supplementary variable.

The concept of typicality of an individual allows to highlight the group of individuals with higher typicality.

Definition 22 *The group of individuals is divided in two parts by using k-means on the typicality variable (i.e. minimising the intra-group variance of typicalities), so that we call optimal group to that whose typicalities are higher.*

Another distance that allows us to define the concept of contribution is:

$$d_2(x, C) := \left(\frac{1}{g} \sum_{i=1}^{g} (1 - \varphi_{x,i})^2 \right)^{1/2}$$

where g are all subclasses of C formed at previous levels (including C).

Definition 23 *We define the contribution of the individual x to the formation of class C at level i of the hierarchical tree as*

$$\gamma_2(x, C) := 1 - d_2(x, C)$$

The notion of contribution is defined to determine the individuals that contribute well to the creation of the rule. These individuals are more responsible than others for forming the rule.

Observation 1 *In the case of the first level of the hierarchy, the formula to calculate the contribution is simplified. If we want to calculate the contribution of any individual to the class formed at the level $i = 1$ of the hierarchy, there is only one subclass of $C = (a, b)$, which is just (a, b). So,*

$$\gamma_2(x, (a, b)) = 1 - d_2(x, (a, b)) = 1 - \left(\frac{1}{1} \sum_{i=1}^{1} (1 - \varphi_{x,1})^2 \right)^{1/2} = 1 - (1 - \varphi_{x,1}) = \varphi_{x,1}$$

Example 8 *Following with Example 1, we will compute the typicality and contribution of the 20 individuals in the construction of the similarity class (HIP,PUN) formed at level 1 of the hierarchy.*

			Contribution	*Typicality*	
x	HIP	PUN	$\varphi_{x,1} = \gamma_2(x,(HIP,PUN))$	$d_1(x,(HIP,PUN))$	$\gamma_1(x,(HIP,PUN))$
$s1$	1	1	1	0.2327	0.94275
$s2$	1	1	1	0.2327	0.94275
$s3$	1	1	1	0.2327	0.94275
$s4$	0	0	0.5	1.91603	0.52862
$s5$	1	1	1	0.2327	0.94275
$s6$	0	0	0.5	1.91603	0.52862
$s7$	1	1	1	0.2327	0.94275
$s8$	0	0	0.5	1.91603	0.52862
$s9$	1	1	1	0.2327	0.94275
$s10$	0	0	0.5	1.91603	0.52862
$s11$	0	1	1	0.2327	0.94275
$s12$	1	1	1	0.2327	0.94275
$s13$	1	1	1	0.2327	0.94275
$s14$	0	0	0.5	1.91603	0.52862
$s15$	0	0	0.5	1.91603	0.52862
$s16$	1	1	1	0.2327	0.94275
$s17$	1	0	0	4.06475	0
$s18$	0	0	0.5	1.91603	0.52862
$s19$	1	1	1	0.2327	0.94275
$s20$	0	0	0.5	1.91603	0.52862

Table 2.15: Typicality and contribution of the individuals to the formation of the similarity class (HIP,PUN).

At the second and third columns of Table 2.15, we show the values of the variables HIP and PUN for each of the students (brought from Table 2.2). Then we compute the contribution of each individual to the class $C = (HIP, PUN)$ (using formula from Observation 1) and we show it at the fourth column. At the fifth and sixth columns we show the distance and typicality calculated as follows.

At Table 2.3, we got the similiarity index between variables HIP and PUN, which form class C: $s(HIP, PUN) = 0.9458524$. So, for example, for the individual s1:

$$d_1(s1,(HIP,PUN)) = \left(\frac{1}{1}\sum_{i=1}^{1}\frac{(\varphi_i - \varphi_{x,i})^2}{1-\varphi_i}\right)^{1/2} = \left(\frac{(0.9458524-1)^2}{1-0.9458524}\right)^{1/2} = 0.2327$$

When we have calculated all distances, then we can calculate the typicalities. Let us calculate the typicality of, for example, s1:

$$\gamma_1(s1,(HIP,PUN)) = 1 - \frac{0.2327}{\max_{x\in I} d_1(x,C)} = 1 - \frac{0.2327}{4.06475} = 0.94275$$

Example 9 *Following with Example 1, now we will calculate the typicality and contribution of the individual s1 to the cohesion class ((HEA,JAZ),REG) formed at level* $i = 3$ *of the hierarchy. The formulae are the same, but now we will consider* φ_i *as the cohesion index instead of the similarity index.*

The cohesion indices $c(HEA, REG) = c(JAZ, REG) = 0.8976726$ *have the same value. So, we will consider that, for example, (HEA,REG) is the generic pair of the class* $C = ((HEA, JAZ), REG)$*. So, attending to Definition 19,* $\varphi_3 = 0.8976726$*. The class* C *has a subclass which is distinct from itself,* $C_1 = (HEA, JAZ)$*. The generic pair of this class* C_1 *formed at level* $i = 1$ *is (HEA,JAZ), whose cohesion index is* $\varphi_1 = c(HEA, JAZ) = 0.9862913$*.*

Now we can calculate the distances, typicality and contribution (shown at Table 2.16). We show the details of the calculations for the individual s1:

$$d_1(s1, C) = \left(\frac{1}{2}\left(\frac{(0.9862913 - 0.5)^2}{1 - 0.9862913} + \frac{(0.8976726 - 0.5)^2}{1 - 0.8976726}\right)\right)^{1/2} = 3.0656$$

$$\gamma_1(s1, C) = 1 - \frac{d_1(s1, C)}{\max_{x \in I} d_1(x, C)} = 1 - \frac{3.0656}{6.2783} = 0.5117$$

$$d_2(s1, C) = \left(\frac{1}{2}\left((1 - 0.5)^2 + (1 - 0.5)^2\right)\right)^{1/2} = 0.5$$

$$\gamma_2(s1, C) = 1 - d_2(s1, C) = 1 - 0.5 = 0.5$$

						Typicality		*Contribution*	
x	JAZ	HEA	REG	$\varphi_{x,1}$	$\varphi_{x,3}$	$d_1(x,C)$	$\gamma_1(x,C)$	$d_2(x,C)$	$\gamma_2(x,C)$
$s1$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s2$	1	1	1	1	1	0.2409	0.9616	0	1
$s3$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s4$	1	1	1	1	1	0.2409	0.9616	0	1
$s5$	1	1	1	1	1	0.2409	0.9616	0	1
$s6$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s7$	1	1	1	1	1	0.2409	0.9616	0	1
$s8$	1	0	0	1	0.5	0.8829	0.8594	0.3536	0.6464
$s9$	0	0	1	0.5	1	2.9456	0.5308	0.3536	0.6464
$s10$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s11$	1	1	1	1	1	0.2409	0.9616	0	1
$s12$	1	1	1	1	1	0.2409	0.9616	0	1
$s13$	0	1	0	0	0	6.2783	0	1	0
$s14$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s15$	1	0	0	1	0.5	0.8829	0.8594	0.3536	0.6464
$s16$	1	1	1	1	1	0.2409	0.9616	0	1
$s17$	0	0	1	0.5	1	2.9456	0.5308	0.3536	0.6464
$s18$	0	0	0	0.5	0.5	3.0656	0.5117	0.5	0.5
$s19$	1	1	1	1	1	0.2409	0.9616	0	1
$s20$	0	1	0	0	0	6.2783	0	1	0

Table 2.16: Typicality and contribution of the individuals to the formation of the cohesion class ((HEA,JAZ),REG).

Chapter 3

The classical Gras implication intensity as a random variable

In this chapter, we review the paper [2]. We analyse the behavior of the implication intensity under sampling of a bivariate binary process. Usually, practitioners use software CHIC in order to get the implicative graph, showing the strongest implications among the variables, and interpreting the important implications in the respective knowledge domain. Our focus is studying the sampling variation of the implication intensity in processes under our control, so that we can decide whether the values found in a sample are reliable or not.

We restrict to the classical version of the implication intensity in the binomial modelisation.

The joint distribution of a binary random variable (X, Y) is completely determined by the joint probability table (completed with the marginal probabilities) shown in Table 3.1.

If we consider the process of sampling from (X, Y) with size n, we can consider the random frequency table given in Table 3.2. The symbol $N_{X\overline{Y}}$ denotes the random number of counterexamples to the rule $X \to Y$, found in the generic sample of size n.

One particular realisation of the random joint frequency table is denoted as shown in Table 3.3. Hence, the symbol $n_{X\overline{Y}}$ denotes the observed number of counterexamples to the rule, found in the particular sample of size n.

The implication intensity $\varphi(X, Y)$ can be seen as a random variable: for each realisation of

		Y		
		0	1	Margin X
X	0	$p_{\overline{XY}}$	$p_{\overline{X}Y}$	$p_{\overline{X}}$
	1	$p_{X\overline{Y}}$	p_{XY}	p_X
	Margin Y	$p_{\overline{Y}}$	p_Y	1

Table 3.1: Joint probability table of (X, Y).

		Y		
		0	1	Margin X
X	0	$N_{\overline{X}\overline{Y}}$	$N_{\overline{X}Y}$	$N_{\overline{X}}$
	1	$N_{X\overline{Y}}$	N_{XY}	N_X
	Margin Y	$N_{\overline{Y}}$	N_Y	n

Table 3.2: Random joint frequency table for generic samples of size n of (X, Y).

		Y		
		0	1	Margin X
X	0	$n_{\overline{X}\overline{Y}}$	$n_{\overline{X}Y}$	$n_{\overline{X}}$
	1	$n_{X\overline{Y}}$	n_{XY}	n_X
	Margin Y	$n_{\overline{Y}}$	n_Y	n

Table 3.3: Particular realisation of the sample of size n of (X, Y).

(X, Y) a value of $\varphi(X, Y)$ is obtained, as a function of the number of successes and the number of counterexamples to the rule $X \to Y$ observed in the sample.

We will present the formula of the probability function of $\varphi(X, Y)$ and its expectation for general sample size and marginal and joint success probabilities. The importance of this result lies in the consideration of $\varphi(X, Y)$ as a populational statistic, and not only as a sample statistic.

Our modelisation of the random binary variables, with a fixed sample size n, leads us to use the binomial model for $N_{X\overline{Y}}$, i.e. $N_{X\overline{Y}} \sim \text{Bin}\left(n, \frac{n_X n_{\overline{Y}}}{n^2}\right)$.

Therefore, the four random variables in the random joint frequency table shown in Table 3.2 form a random vector which follows a multinomial distribution: $(N_{\overline{X}\overline{Y}}, N_{\overline{X}Y}, N_{X\overline{Y}}, N_{XY}) \sim M_4(n, p_{\overline{X}\overline{Y}}, p_{\overline{X}Y}, p_{X\overline{Y}}, p_{XY})$. Consequently, $\varphi(X, Y)$ varies at every sample (of size n) from (X, Y), thus it is a random variable.

Here we derive the general formula of the probability funcion of $\varphi(X, Y)$ and of its expectation.

On the one hand, and conditioned to a sample of size n, the probability function for the vector of absolute frequencies is:

$$P(N_{\overline{X}\overline{Y}} = n_{\overline{X}\overline{Y}}, N_{\overline{X}Y} = n_{\overline{X}Y}, N_{X\overline{Y}} = n_{X\overline{Y}}, N_{XY} = n_{XY}) =$$
$$= \frac{n!}{n_{\overline{X}\overline{Y}}! n_{\overline{X}Y}! n_{X\overline{Y}}! n_{XY}!} p_{\overline{X}\overline{Y}}^{n_{\overline{X}\overline{Y}}} p_{\overline{X}Y}^{n_{\overline{X}Y}} p_{X\overline{Y}}^{n_{X\overline{Y}}} p_{XY}^{n_{XY}}$$

The value of $\varphi(X, Y)$ conditioned to $(N_{\overline{X}\overline{Y}} = n_{\overline{X}\overline{Y}}, N_{\overline{X}Y} = n_{\overline{X}Y}, N_{X\overline{Y}} = n_{X\overline{Y}}, N_{XY} = n_{XY})$ is:

$$\varphi(X, Y) = 1 - F_{N_{X\overline{Y}}}(n_{X\overline{Y}})$$

where $F_{N_{X\overline{Y}}}$ represents the cumulative distribution function of the variable $N_{X\overline{Y}}$. Hence, the probability function for the random variable $\varphi(X,Y)$ can be written as:

$$f(\varphi_0) := P(\varphi(X,Y) = \varphi_0) = \sum_{\substack{n_{\overline{XY}}+n_{\overline{X}Y}+ \\ n_{X\overline{Y}}+n_{XY}=n}} \frac{n!}{n_{\overline{XY}}!n_{\overline{X}Y}!n_{X\overline{Y}}!n_{XY}!} p_{\overline{XY}}^{n_{\overline{XY}}} p_{\overline{X}Y}^{n_{\overline{X}Y}} p_{X\overline{Y}}^{n_{X\overline{Y}}} p_{XY}^{n_{XY}}$$

where the summation corresponds to every vector $(n_{\overline{XY}}, n_{\overline{X}Y}, n_{X\overline{Y}}, n_{XY})$ of nonnegative integers such that $\varphi_0 = 1 - F_{N_{X\overline{Y}}}(n_{X\overline{Y}})$ and such that $n_{\overline{XY}} + n_{\overline{X}Y} + n_{X\overline{Y}} + n_{XY} = n$.

For the expectation of $\varphi(X,Y)$, we can use the expression of $\varphi(X,Y)$ conditioned to values $(n_{\overline{XY}}, n_{\overline{X}Y}, n_{X\overline{Y}}, n_{XY})$, and the definition to get:

$$\begin{array}{l} E(\varphi(X,Y)) = \\ = \sum_{\substack{n_{\overline{XY}}+n_{\overline{X}Y}+ \\ n_{X\overline{Y}}+n_{XY}=n}} \left(1 - F_{N_{X\overline{Y}}}(n_{X\overline{Y}})\right) \times \frac{n!}{n_{\overline{XY}}!n_{\overline{X}Y}!n_{X\overline{Y}}!n_{XY}!} p_{\overline{XY}}^{n_{\overline{XY}}} p_{\overline{X}Y}^{n_{\overline{X}Y}} p_{X\overline{Y}}^{n_{X\overline{Y}}} p_{XY}^{n_{XY}} \end{array}$$

where the summation corresponds to every vector $(n_{\overline{XY}}, n_{\overline{X}Y}, n_{X\overline{Y}}, n_{XY})$ of nonnegative integers such that $n_{\overline{XY}} + n_{\overline{X}Y} + n_{X\overline{Y}} + n_{XY} = n$.

If the marginal probabilities p_X and p_Y and the sample size n are to be fixed, then we can see the effect of the parameter $p_{Y|X}$ (the theoretical confidence of the rule $X \to Y$) on the complete distribution of $\varphi(X,Y)$ and on its expected value $E(\varphi(X,Y))$.

For instance, Figure 3.1 shows the distribution of $\varphi(X,Y)$ for fixed values $p_X = 0.5$, $p_Y = 0.5$, and $n = 30$, and different values of $P(Y|X)$. Then we show the effect of this parameter on the probability function of $\varphi(X,Y)$.

Figure 3.2 shows the effect of the conditional probability $p_{Y|X}$ on the mean value of $\varphi(X,Y)$ for fixed $p_X = 0.5$, $p_Y = 0.5$, and $n = 30$.

We provide the R scripts in the Appendix A. A more complete study is shown in [2].

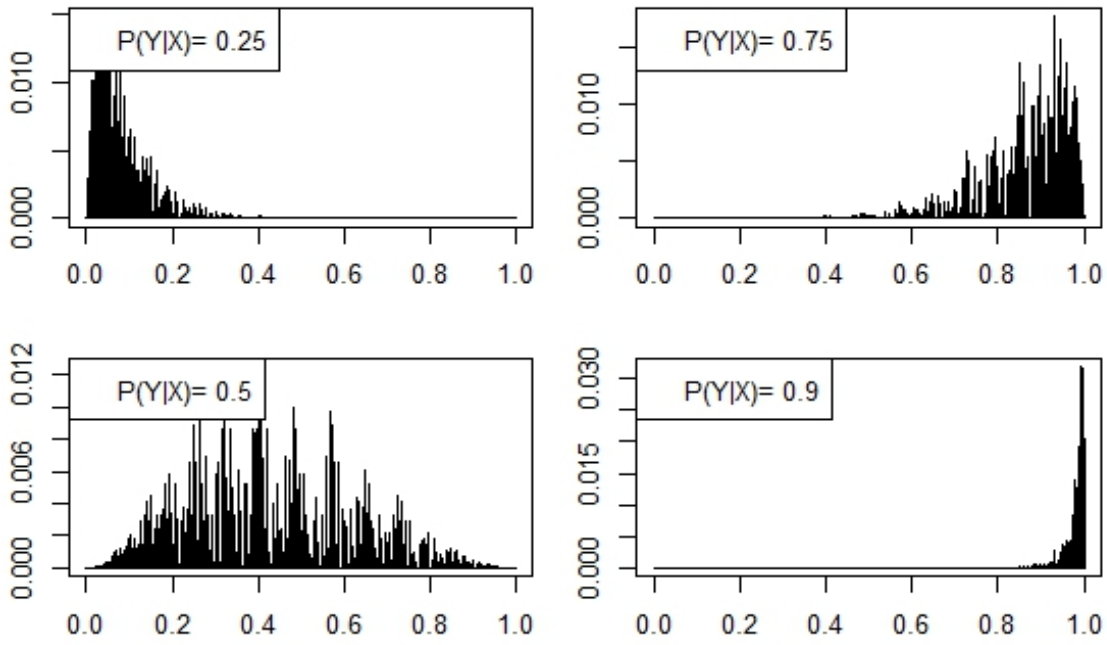

Figure 3.1: Density function of $\varphi(X, Y)$ for different values of $P(Y|X)$.

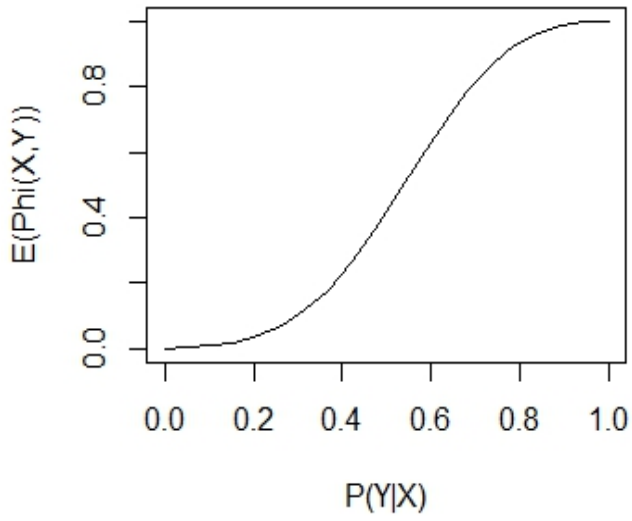

Figure 3.2: Expectation of $\varphi(X, Y)$.

Chapter 4

Conclusions

As we have seen along this work, SIA bases its results by contrasting the observed sample with what the pure randomness (independence) would produce, measuring the discard of one respect to the other. It provides tools to discover relationships of different types between variables (implications, similarity, cohesion, rules of rules, typicality, contribution...).

Regarding its applications, it is very suitable for sciences such as didactics, sociology, psychology, where complex situations are analyzed and experimentation gives rise to a large number of variables.

It is an area with many things to explore still, and that can be very helpful for many areas of society.

Appendix A

Here is the R code to produce Figures 3.1 and 3.2.

Listing A.1: Code for computing the density of φ, and plotting density and expectation

```
# COMPUTATION OF PHI

phi0 = function(x){
  # computes the value of phi0 for a particular sample
  # with x[1] in nXnY, x[2] in nXY, x[3] in XnY and x[4] in XY
  return(1-pbinom(q=x[3], size=sum(x),
                  prob=((x[3]+x[4])*(x[1]+x[3]))/((sum(x))^2)))
}

# COMPUTATION OF DENSITY AND EXPECTATION OF PHI

rvgrasphi = function(pX=0.5, pY=0.5, pXY=NULL, pYgivenX=NULL, n=10){
  # probability function and expectation for the
  # Gras implication index of two Bernoulli variables
  # X and Y of parameters pX and pY and joint success
  # probability pXY (or conditional probability pYgivenX).
  # It returns a list of two components:
  # $f = the values of phi and their probability
  # $E = the expected value
  require(partitions) #needs the package
  if( is.null(pXY) ){
    pXY = pX * pYgivenX
  }else{
    pYgivenX = pXY/pX
  }
  pnXnY = 1 - pX - pY + pXY
  pnXY = pY - pXY
  pXnY = pX - pXY
  # PROBABILITY FUNCTION FORMULA f(x) := Pr(Phi=x)
  # f(x) = sum_[nn : phi(nn)=x] prob(NN=nn)
  # where nn are all the possible 4 joint absoulte frequencies
  # Compute all phi.nn, and sum probabilities of repeated values
  nn = compositions(n,4)
```

```
  # computation of prob(NN=nn)
  pr.nn = apply(X=nn, MAR=2, FUN='dmultinom', size=n,
                prob=c(pnXnY, pnXY, pXnY, pXY))
  # computation of phi(nn)
  phi.nn = apply(X=nn, MAR=2, FUN='phi0')
  phi.values = sort(phi.nn)[c((1:(length(phi.nn)-1))
                              [as.logical(sign(diff(sort(phi.nn))))],
                              length(phi.nn))]
  phi.prob = diff(c(0,cumsum(pr.nn[order(phi.nn)])
                    [c((1:(length(phi.nn)-1))
                       [as.logical(sign(diff(sort(phi.nn))))],
                       length(phi.nn))]))
  Ephi = sum(phi.nn * pr.nn)
  result = list( f=data.frame(phi=phi.values, fphi=phi.prob),
                 E=Ephi )
  return( result )
}

# PLOTTING DENSITY OF PHI

pYgivenX=c(0.25, 0.50, 0.75, 0.90)
par(mar=c(3,2,1,1)+0.1)
layout(matrix(1:4,2,2))
for(i in 1:4){
  result=rvgrasphi(pX=0.5, pY=0.5, pXY=NULL, pYgivenX=pYgivenX[i],
                   n=30)
  plot(x=result$f$phi, y=result$f$fphi, type='h')
  legend(x='topleft', legend=paste('P(Y|X)=', pYgivenX[i],
                                   collapse=''))
}

# PLOTTING EXPECTATION OF PHI

# range of P(Y|X)
pX=0.5
pY=0.5
n=30
pXYmin = max(c((pX + pY)-1, 0))
pXYmax = min(c(pX,pY))
pYgivenXmin = pXYmin/pX
pYgivenXmax = pXYmax/pX
p = seq(fr=pYgivenXmin, to=pYgivenXmax, len=20)
Ephi = numeric(0)
for( pYgivenX in p){
  pXY = pX * pYgivenX
  pnXnY = 1 - (pX + pY) + pXY
  pnXY = pY - pXY
  pXnY = pX - pXY
  nn = compositions(n,4)
```

```
  # prob(NN=nn)
  pr.nn = apply(X=nn, MAR=2, FUN='dmultinom', size=n,
                prob=c(pnXnY, pnXY, pXnY, pXY))
  # phi(nn)
  phi.nn = apply(X=nn, MAR=2, FUN='phi0')
  Ephi = c(Ephi, sum(phi.nn * pr.nn))
}
layout(matrix(1:1,1,1))
par(mar=c(5, 4, 4, 2) + 0.1)
plot(x=p, y=Ephi, type='l', ylim=c(0,1), xlab='P(Y|X)',
     ylab='E(Phi(X,Y))')
```

Bibliography

[1] Antoine Bodin. Analyse implicative: modèles sous-jacents à l'analyse implicative et outils complémentaires. *IRMAR*, 97(32), 1997.

[2] Pablo Gregori Huerta, Raphaël Couturier, and Rubén Pazmiño Mají. On the probability distribution of the classical gras implication index between two binary random variables. In *VII International Conference on Statistical Implicative Analysis*, Brazil, November 2013.

[3] Josep Vicent Felip i Bardoll. Análisis estadístico implicativo de un estudio sociológico de la localidad de alcublas (valencia). Master's thesis, Universitat Jaume I, Castellón de la Plana, September 2011.

[4] I.C. Lerman. Sur l'analyse des donées préalable à une classification automatique (proposition d'une nouvelle mesure de similarité). *Mathématiques et Sciences Humaines*, (32):5–15, 1970.

[5] Pilar Orús, Larisa Zamora, and Pablo Gregori, editors. *Teoría y aplicaciones del análisis estadístico implicativo: primera aproximación en lengua hispana.* Universitat Jaume I, Departamento de Matemáticas and Universidad de Oriente de Santiago de Cuba, Facultad de Matemática y Computación, Castellón and Santiago de Cuba, 2009.

[6] Irene Pitarch. Estudio sobre la viabilidad y el interés didáctico del tratamiento de la información en la ESO. Master's thesis, Universitat Jaume I, Castellón de la Plana, 2002.